FEBS Letters 580 (2006) 6361–6365

Intron exclusion and the mystery of intron loss

Kejin Hu*,[1]

Pharmacology Department, University of Pittsburgh, W1301, BST, 200 Lothrop Street, PA 15213, USA

Received 17 August 2006; accepted 23 October 2006

Available online 7 November 2006

Edited by Takashi Gojobori

Abstract **Mechanisms for loss and gain of introns are elusive. Reported here is a new pattern of intron loss which features a random loss of a single intron in a multiple-intron gene with its neighboring introns remained, which process is defined as intron exclusion. Intron exclusion is reminiscent of removal of a limited stretch of non-homologous sequence in a homologous recombination (HR) triggered by a double strand break (DSB), and therefore lends further evidence for a theory of intron loss through HR between a cDNA and its genomic intron-containing locus. Thus, a model for intron loss is formulated.**

Keywords: Intron loss; Intron exclusion; Double strand break; Homologous recombination; Alcohol dehydrogenase; Genome streamlining

1. Introduction

Even 28 years after its discovery, many fundamental questions about introns remain unanswered. The long-lasting debate about evolution of introns reaches little agreement between proponents of the competing 'introns-early' and 'introns-late' theories [1–6]. Whether or not introns existed before the divergence of prokaryote and eukaryote, it is certain that there have been losses and gains of introns afterward. However, mechanisms for gain and loss of introns are quite elusive. Five speculative mechanisms are currently proposed for intron gain [7]. But, there is sparse discussion about a mechanism of intron loss. One speculation is the simple genomic deletion of an intron [5]. Lewin [8] first had the idea that gene conversion might occur between a cDNA and its intron-containing genomic locus, and this might have plucked out one of the two introns in the rat one-intron insulin gene. Fink [9] observed that the yeast genome is predominantly intronless, and has a 5′-biased distribution of introns for the rare intron-containing genes. He therefore surmised that homologous recombination (HR) between a full-length or a truncated cDNA and its genomic counterpart is responsible for intron loss in yeast. Elimination of an intron through an RNA-mediated process was later experimentally demonstrated in yeast [10]. Recently, Hu and Leung [11] reported that multiple introns in several *CatL*-like genes from plant and animal are independently eliminated probably in situ without any change to the gene. This observation added support for a mechanism of RNA-mediated HR for intron loss in an active intronless gene. Although HR is the most cited theory about intron loss from an active gene, the underlying mechanism for this process is unclear. We recently reasoned that DNA double-strand-break-repair (DSBR) machinery might be involved in this process, and referred to the pathway as cDNA-mediated homologous recombination with the involvement of DSBR machinery (cDMHR/DSBR) [11]. But, the involvement of DSBR machinery in cDNA-mediated HR is lacking evidence. To better understand intron loss, I employed a simple method to align introns in order to study the intron pattern in plant *Adh* (alcohol dehydrogenase) genes because previous report showed that the 9-intron structure of grass *Adh* is highly conserved [12]. This investigation led to the identification of a new mode of intron loss that implicates a mechanism for intron elimination.

2. Materials and methods

A novel and simple approach was employed to realize the alignment of introns. An intron is regarded as a special codon that encodes an imaginary amino acid X. In the amino acid sequences to be aligned, X (intron) is placed in between the two residues when a phase-0 intron is located in between their corresponding codons; X is placed before the amino acid when the phase-1 intron breaks the corresponding codon; and after the amino acid when a phase-2 intron breaks its codon. This transformation of introns allows for easy and quick alignment of introns in the plant *Adh* genes with the publicly available alignment programs such as the Multiple Sequence Alignment on BCM Search Launcher (http://searchlauncher.bcm.tmc.edu/). The aligned sequences were then highlighted with the Boxshade tool (http://www.ch.embnet.org/software/BOX_form.html).

Comparison of gene structures was restrained to plant *Adh* genes because slippage, loss and gain of introns have occurred during evolution and that make it difficult to compare introns among distant species, for example, introns of human *Adh5* (NP_000662) and *Arabidopsis Adh* (NP_564409) are not alignable although their product share 47% aa identity. Intron dynamics can be more confidently uncovered by comparison of gene structures from closely related species with one from a relatively distant species as an outgroup [11]. It was reported that plant *Adh* genes have highly conserved structures, and this permitted unambiguous identification of intron loss in plant *Adh* genes in this study.

3. Results and discussion

The alignment in Fig. 1 shows that the 9-intron structure of *Adh* is much more widely conserved than previously reported.

*Fax: +1 412 648 1945.
E-mail address: hukejin@gmail.com

[1]Present address: Department of Pharmacology, University of Wisconsin, Room 2683, Medical Science Center, 1300 University Avenue, Madison, WI 53706, USA. Fax: +1 608 262 1257.

Abbreviations: ADH, Alcohol dehydrogenase; cDMHR, cDNA-mediated homologous recombination; DSB, Double strand break; DSBR, DSB repair; HR, Homologous recombination

0014-5793/$32.00
doi:10.1016/j.febslet.2006.10.048

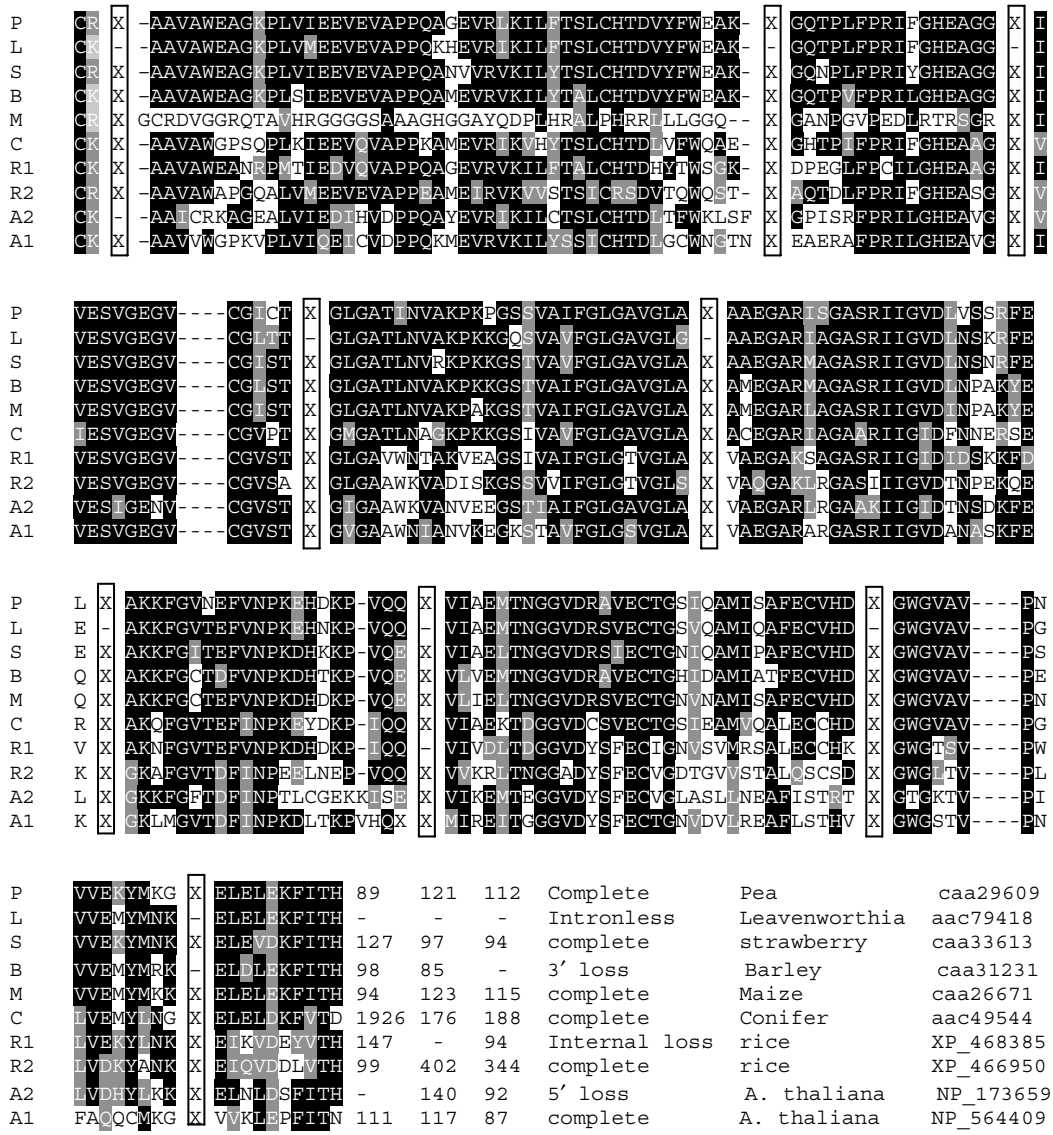

Fig. 1. Alignment of introns for selected plant Adh, showing intron exclusion. Boxed are introns. X, presence of intron, – inside boxes indicates absence of intron. First column in each panel is symbol of organism name which are initials of plant names followed by numeral if two homologs are presented for one organism. The matrix after the alignment shows the intron length in bp for intron 1, 7 and 9 which were found excluded in this study. After intron length matrix are description of intron pattern, plant names and accession numbers. Intron alignment was achieved simply by treating an intron as a codon encoding an 'amino acid' 'X'. For compactness, the N termini and C termini were not shown here, neither the two fragments between intron 3 and 4, and between intron 8 and 9 where a ---- indicates this fact.

This same structure is found from angiosperm to gymnosperms, and both monocot and dicot *Adh* identically have this 9-intron structure. It can be seen that aac79418 (*Adh-3*) has lost all of the 9 introns. Several lines of evidence suggest that *Adh-3* is active [13]. This is in agreement with our previous suggestion that HR between full length cDNA and its genomic locus has replaced the intron-containing genomic copy, and therefore removes all of the 9 introns. The HR theory of intron loss also predicts that HR with the involvement of a truncated cDNA results in biased distribution of introns. However, apart from the complete, precise and innocuous intron loss in *Adh-3*, I found a new pattern of intron loss which is hard to be explained by truncation hypothesis. It is found that the first intron of NP_173659 is lost. The last intron in CAA31231 is absent as well. A single intron in the middle of XP_468385 is missing with the surrounding introns retained. All of these three cases are precise in-frame losses of a single intron. For NP_173659 and CAA31231, there are 5′- and 3′- biased intron loss respectively rather than 3′- and 5′-biased retention of introns that is expected by truncated-cDNA assumption. Furthermore, the location of the singly lost intron appears to be random.

Is the random loss of a single intron (hereafter referred to as intron exclusion) in the multiple-intron *Adh* genes in conflict with cDMHR/DSBR theory? This is what confused some authors when they found that single intron deletion are more frequent than simultaneous losses of several introns [5,14]. I believe intron exclusion does not conflict with HR theory of intron loss. On the contrary, this mode of intron loss adds compelling evidence to our cDMHR/DSBR model. If we know the

nature of DSB repair, truncation theory is totally unnecessary in explaining intron exclusion. One key component of our model is the involvement of DSBR machinery. Assuming that a DSB occurs at either junctions of an intron, a gap repair process through an HR between a broken DNA and its cDNA will remove not all the introns but the single intron where DSB occurs. This mode of crossover was demonstrated in yeast [15]. They showed that crossover occurred at the closest homologous regions flanking the break. Even though the DSB occurs within the intron, the broken intron can be eliminated since experimental evidence proved that a limited stretch of non-homologous DNA is permitted on both ends of the break [15]. In fact, gap repair has become a routine technique in yeast on many occasions such as allele recovery/transfer, gene transfer from plasmid to plasmid, and introducing PCR-based point mutation into a cloned gene.

Many reported losses of introns are in fact intron exclusion. Genome-wide comparison of intron positions revealed five exact losses of a single internal intron in five genes of mouse, and one such in-frame loss of an internal intron in rat [16]. Intron Y, which should be located between the intron 7 and 8, was reported to be lost precisely from *Dfak* of *Drosophila melanogaster* group [17]. Feiber et al. [18] showed that there is an intron presence-absence polymorphism for the *4f-rnp* gene among *Drosophila robusta* population in which intron 7 is deleted precisely with the upstream and downstream introns intact.

In Fig. 2, I propose a model of intron loss. A full-length cDNA can stimulate HR through double crossover events at its two ends since the free ends resemble a DSB which serves as a trigger for DSB repair machinery, and results in loss of all introns. Yeast provides the clearest example for that free homologous ends can promote HR. A standard procedure for yeast gene deletion is that target homologous sequence is attached to both ends of a marker gene (for example URA3) for which the corresponding genomic locus contains a point mutation that inactivates the gene. In this case, homologous recombination occurs between free ends sequence and its genomic locus, not between the internally located marker and its genomic locus (Fig. 2). 5′- or 3′- truncated cDNA permits a biased retention of introns after HR. A break in a specific intron will trigger the removal of this very intron when HR occurs. HR repair of DNA DSB is ubiquitous. It is well-known that DSB promotes HR in yeast and trypanosome. That is true in the more complex genomes. For example, a DSB introduced by P-element in *Drosophila* stimulates HR [19]. Also, HR was triggered by a DSB in mouse initiated by the rare-cutting endonuclease I-SceI introduced from yeast [20,21]. Furthermore, RecA protein, that is a key player in HR DSB repair, is conserved from virus to human. This is consistent with the ubiquitous loss of introns suggested by the current data.

The DSB-trigger theory suggests that yeast and bacterial genomes might have experienced massive streamlining by

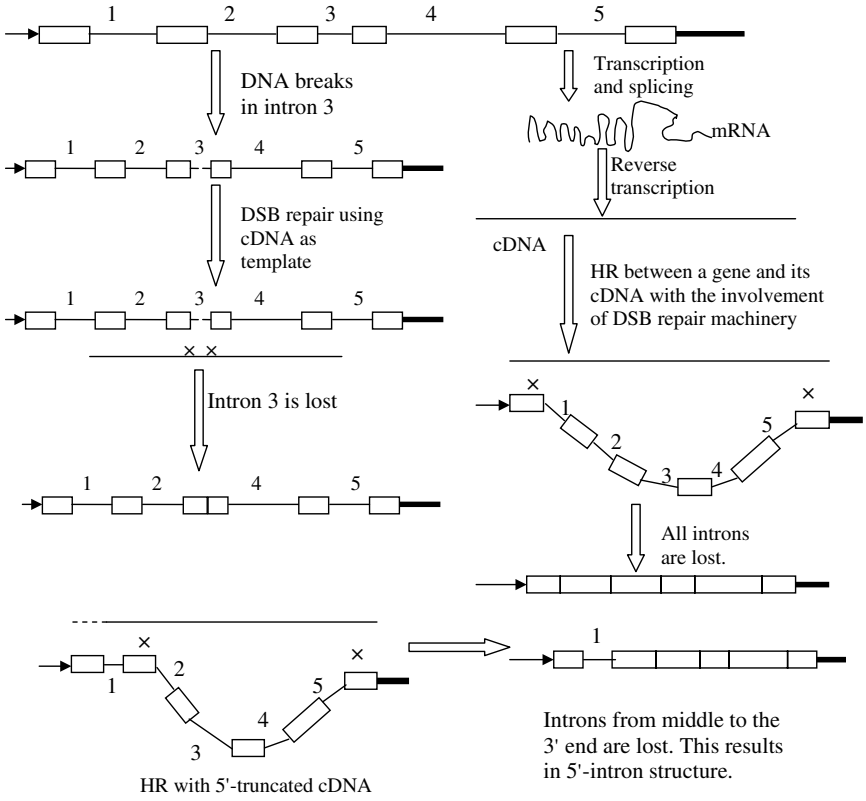

Fig. 2. A model for intron loss. Homologous recombination between cDNA and its genomic intron-containing locus results in loss of introns. Left on upper panel: A DSB within a specific intron stimulates HR, and causes loss of the specific intron. Right on upper panel: When no DSB occurs on genomic locus, free ends of cDNA can still promote cross over, but at both ends of cDNA with its chromosomal loci, and this causes loss of all introns. Lower panel: When a truncated cDNA is involved, intron loss occurs only in the undamaged region, and introns in the region corresponding to the truncated portion are retained. Free thin line is cDNA. Broken line at the beginning of cDNA indicates the truncated portion. Open box denotes exon. Connecting thin line stands for intron. Intron number is given above each intron. A gap in intron 3 designates the double strand break of DNA. Horizontal arrow connected to the first exon is the upstream promoter sequence, and thick line after the last exon is the downstream sequence. × marks the regions where cross over occurs.

eliminating introns. If the exon shuffling theory is correct, we can even hypothesize that bacteria had spliceosomal introns and these introns have all been eliminated by cDMHR/DSBR pathway. *E. coli* experience 3000–5000 DNA lesions per cell per generation [22], and there is no barrier between DNA and cDNA due to a lack of a nucleus. Even type II introns could be eliminated by cDMHR/DSBR pathway triggered by DSB.

To make sense of my DSB-stimulation hypothesis of intron loss, we need to rule out one possibility that incompletely spliced RNA is reverse transcribed and its cDNA subsequently takes part in gene conversion by HR with the parent gene, which is one of the conjectures by Wada et al. [14]. Several lines of evidence make this route impossible. First of all, unspliced or partially spliced species of RNA are unstable and are in nuclear fraction [23,24]. Secondly, given that RT activity is derived from retroviruses or retroelements such as *Ty1-copia* elements, it is believed to exist only in cytoplasm [25]. Furthermore, no nuclear RT activity is documented. Thirdly, with accumulation of reported retropseduogenes, all are intronless copies [26]. To the best of my knowledge, no intron-containing retropseduogene has been reported. For the three intron-excluded genes in this communication, all are active because multiple EST and cDNA entries exist in database for NP_173659 and XP_468385. The intron-excluded barley gene is also expressed [27]. If a cDNA species of an incompletely spliced RNA can replace its genomic locus, it can be transposed more frequently. Current data suggest that this is not the case. Another apparent evidence against a mechanism in which a cDNA of a partially spliced mRNA is involved is that there is no report of a intron-containing cDNA. Genomic deletion was suggested to be a mechanism of intron elimination [5,28]. However, this is an imprecise process as exemplified by *jinwei* in the *Drosophila* population [28]. This imprecise deletion of intron can be detrimental and could not be a general mechanism of precise loss of intron.

One phenomenon that embarrassed intron-early advocates is that some conserved introns are absent from the overwhelming majority of the homologous genes. Intro-late supporters suggest this could be better explained by parallel intron insertion rather than intron loss in a statistical view. cDMHR/DSBR theory of intron loss implies that intron loss is easier and neutral. On the other hand, intron insertion is regarded as a mutagenic process. Therefore, the overwhelming absence of a specific shared intron from orthologous genes can still be explained by intron loss through a cDMHR/DSBR process.

A key finding in favor of the introns-early theory is that the *GAPDH* genes with eubacterial/plastid ancestry and the cytoplasmic *GAPDH*s share five spliceosomal introns [29,30]. However, the assumption that the five shared introns are homologous requires that all introns in eubacteria and most introns in protists were lost independently in a relatively recent time [6]. Massive loss of introns in bacteria, eubacteria and protists were thought unreasonable by the intron late proponents [4,6]. cDMHR/DSBR theory explains to us how the massive loss of introns in bacteria, eubacteria and protist is readily accomplished. Unicellular and fast-growing organisms are under the pressure of selection for genome streamlining for the sake of cell economy in DNA replication, transcription, splicing and enzymic degradation of spliced introns [11,31]. Both intergenic and intronic sequences are the target for genomic streamlining. An appealing scenario is that DSB dramatically sped up the genomic streamlining in intervening sequence in the currently intronless and intron-poor genomes by promoting HR between intron-containing genomic copy and its cDNA.

Given that DSB triggers HR between cDNA and its genomic locus, why do some genomes retain more introns than others? Several factors might account for these differences. First, intron length may contribute. Long intron might prevent such an HR. Interestingly, the reported intron exclusion all occur with short intron [16–18]. Interestingly, the introns that are found excluded in this study are generally short except for conifer *Adh* intron 1 (Fig. 1). Second, intron density might be a factor. For example, the human genome has more introns per gene and its intron is generally longer. A high density of intron in a gene might reduce HR between the cDNA and its genomic copy. Third, the different mechanisms of DSB repair in different genomes might be a factor also. For example, yeast uses HR as a default mechanism to repair DSB while vertebrates use non-homologous end joining (NHEJ) more frequently to join the broken ends [32]. There might be other unknown factors that contribute to various intron profiles in different genomes. For example, abundance of the activity of reverse transcriptase, stability of a specific mRNA, presence/absence of nuclear envelope and presence/absence of germ line might all result in differing intron profiles. Whether these factors are contributing to the different intron profile in different genomes awaits further investigation.

4. Summary

In this study, with a simple and improved approach of intron alignment, I unambiguously identified and defined a new mode of intron loss: intron exclusion which is a process of the precise removal of a single intron from a multiple-intron gene with the surrounding introns remained. This new mode of intron loss implicates that DSB might be a trigger of precise intron loss. A model for intron loss is proposed, that can accommodate all intron data available. This model can explain simultaneous intron loss, intron exclusion and biased-intron retention or biased intron loss.

Acknowledgements: I thank professor Kenneth Kemphues from Cornell University for his critical reading of my manuscript. I am in debt to Dr. Marjet Heitzer, from University of Pittsburgh for her discussions on this manuscript.

References

[1] Roy, S.W. (2003) Recent evidence for exon theory of genes. Genetica 118, 251–266.
[2] Lynch, M. (2002) Intron evolution as a population-genetic process. PNAS 99, 6118–6123.
[3] De Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S. and Gilbert, W. (1998) Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. PNAS 95, 5094.
[4] Logsdon, J.M. (1998) The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. 8, 637–648.
[5] Lynch, M. and Richardson, A.O. (2002) The evolution of spliceosomal introns. Curr. Opin. Genet. Dev. 12, 701–710.
[6] Logsdon, J.M., Palmer, J.D., Stoltzfus, A. and Cerff, R. (1994) Origin of introns – early or late? Nature 369, 526–528.

[7] Coghlan, A. and Wolfe, K.H. (2004) Origins of recently gained introns in Caenorhabditis. PNAS 101, 11362–11367.
[8] Lewin, R. (1983) How mammalian RNA returns to its genome. Science 219, 1052–1054.
[9] Fink, G.R. (1987) Pseudogenes in yeast? Cell 49, 5–6.
[10] Derr, L.K. and Strathern, J.N. (1993) A role for reverse transcripts in gene conversion. Nature 361, 170–173.
[11] Hu, K.J. and Leung, P.C. (2006) Complete, precise, and innocuous loss of multiple introns in the currently intronless, active cathepsin L-like genes, and inference from this event. Mol. Phylogenet. Evol. 38, 685–696.
[12] Gaut, B.S., Peek, A.S., Morton, B.R. and Clegg, M.T. (1999) Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaccae). Mol. Biol. Evol. 16, 1086–1097.
[13] Charlesworth, D., Liu, F.L. and Zhang, L. (1998) The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus Leavenworthia (Brassicaceae). Mol. Biol. Evol. 15, 552–559.
[14] Wada, H., Kobayashi, M., Sato, R., Satoh, N., Miyasaka, H. and Shirayama, Y. (2002) Dynamic insertion-deletion of introns in Deuterostome EF-1α genes. J. Mol. Evol. 54, 118–128.
[15] Ma, H., Kunes, S., Schatz, P.J. and Botstein, D. (1987) Plasmid construction by homologous recombination in yeast. Gene 58, 201–216.
[16] Roy, S.W., Fedorov, A. and Gilbert, W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. PNAS 100, 7158–7162.
[17] Jin, S., Hu, G.A., Qian, Y.H., Zhang, L., Zhang, J., Qiu, G., Zeng, Q.T. and Gui, J.F. (2005) Identification of one intron loss and phylogenetic evolution of Dfak gene in the Drosophila melanogaster species group. Genetica 125, 223–230.
[18] Feiber, A.L., Rangarajan, J. and Vaughn, J.C. (2002) The evolution of single-copy Drosophila nuclear 4f-rnp genes: spliceosomal intron losses create polymorphic alleles. J. Mol. Evol. 55, 401–413.
[19] Keeler, K.J., Dray, T., Penney, J.E. and Gloor, G.B. (1996) Gene targeting of plasmid-borne sequence to a double-strand DNA break in Drosophila melanogaster. Mol. Cell. Biol. 16, 522–528.
[20] Rouet, P., Smih, F. and Jasin, M. (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol. Cell. Biol. 14, 8096–8106.
[21] Choulika, A., Perrin, A., Dujon, B. and Nicolas, J. (1995) Introduction of homologous recombination in mammalian chromosomes by using the I-sceI system of *Saccharomyces cerevisiae*. Mol. Cell. Biol. 15, 1968–1973.
[22] Cox, M. (1999) Recombinational DNA repair in bacteria and the RecA protein. Prog. Nucleic Acid Res. Mol. Biol. 63, 311–366.
[23] Lewin, B. (2000) Nuclear splicingGene VII, pp. 659–718, Oxford University Press, New York.
[24] Luo, M. and Reed, R. (1999) Splicing is required for rapid and efficient mRNA export in metazoans. PNAS 96, 14937–14942.
[25] Wilhelm, F.X., Wilhelm, M. and Gabriel, A. (2005) Reverse transcriptase and integrase of the Saccharomyces cerevisiae Ty1 element. Cytogenet. Genome Res. 110, 269–287.
[26] Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986) Norviral retroposons: genes, pseudogenes, and transposable elements generated by reverse flow of genetic information. Ann. Rev. Biochem. 55, 631–661.
[27] Trick, M., Dennies, E.S., Edwards, K.J.R. and Peacock, W.J. (1988) Molecular analysis of the alcohol dehydrogenase gene family of barley. Plant Mol. Biol. 11, 147–160.
[28] Liopart, A., Comeron, J.M., Brunet, F.G., Lachaise, D. and Long, M. (2002) Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection. PNAS 99, 8121–8126.
[29] Quigley, F., Martin, W.F. and Cerff, R. (1988) Intron conservation across the prokaryote-eukaryote boundary: structure of the nuclear gene for chloroplast glyceraldehydes-3-phosphate dehydrogenase from maize. PNAS 85, 2672–2676.
[30] Kersanach, R., Brinkmann, H., Liaud, M., Zhang, D., Martin, W. and Cerff, R. (1994) Five identical intron positions in ancient duplicated genes of eubacterial origin. Nature 367, 387–389.
[31] Cavalier-smith, T. (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. 34, 247–278.
[32] Aylon, Y. and Kupiec, M. (2004) DSB repair: the yeast paradigm. DNA repair 3, 797–815.